

Mentoring a New Generation of Data Scientists

With massive amounts of data produced at LLNL, data science is becoming a vitally important field, allowing Laboratory programs to gain new insights into mission objectives and make better decisions. The [Data Heroes summer intern program](#) began in 2015 as a collaboration between the Engineering and Computation directorates to provide hands-on training and real-world experience for the next generation of LLNL data scientists. “The Laboratory has interesting problems to solve in this field,” notes Livermore statistician Ana Kupresanin.

Though only in its second year, the 2016 Data Heroes program drew hundreds of applicants. Directors Ghaleb Abdulla, Jim Gansemer, and Kelley Herndon Ford invited 20 graduates and undergraduates to work on research projects proposed by LLNL scientists. The program supplements interns’ education with seminars, courses, and facility tours, and Data Heroes cap off their internships by participating in the Student Poster Symposium. Staff from both directorates mentor students through all aspects of the program.

Students interested in data science have many opportunities in the Bay Area, so flexibility is crucial to the program’s recruiting strategy. Internships do not have to begin at the same time, and terms vary from 10 to 12 weeks. Data Heroes are even allowed to come back for another year. Moreover, working on funded projects means using real data with access to top-level scientists and facilities. Tim Moon, a Stanford University Ph.D. student, had not worked with supercomputers until he came to the Laboratory, which allowed him to “get very solid footing” in computational mathematics. Similarly, Olivia Simpson, a Ph.D. candidate at the University of California, San Diego, was able to use LLNL’s in-house software Havoq-GT to analyze computer network changes over time, while the program exposed Harvard University’s Elizabeth Tran to computer science for the first time.

Data Heroes spend half their time on their mentors’ projects. Kupresanin structures her research so interns can gain a sense of accomplishment, and emphasizes the importance of practical learning beyond academia: “Students work on data they wouldn’t see in commercial or academic environments,” she points out. “Our projects are not homework problems with solutions in the back of the book.” The 2016 class included Brendan McVeigh, who worked with mentor Alan Kaplan and a University of California, San Francisco, neuroscience team to analyze electrocorticographic medical data and understand physical properties of the brain. Data Hero Eliza Wells explored efficiency differences between liquid and air cooling at Computation facilities. Paul Rosenfield, a Data Hero from Stanford University, praised the collaboration between physicists and statisticians on his neutron detection project. “It was a unique opportunity to be in as a statistician,” he explains, because the team was able to generate data as they went and develop their detector iteratively.

The Data Hero internship also provides a series of courses taught by LLNL staff; students select three to four courses aligned with their projects and interests. This year the program offered nine courses across a range of topics such as Bayesian methods, applied predictive modeling, and various open-source software and programming languages. Lawrence Scholar and Data Hero Devin Francom describes the curriculum as cutting edge: “The Laboratory is just starting to use TensorFlow and is already teaching it to us,” he says, referring to software that enables machine learning like pattern recognition.

Kupresanin's experience interning at two national laboratories compelled her to mentor Data Heroes, and she enjoys learning from the younger generation. "They might look at a problem in a different way or ask questions I might not think of," she says. "They bring a different energy to my work." Students, too, leverage this relationship. For example, Francom has spent three summers at LLNL—two as a Data Hero—working with his mentors on Bayesian computational models.

As a young program, the Data Heroes internship will evolve. According to Gansemer, future summers could focus on a single topic or expand course offerings, and he and Kupresanin hope the program continues to attract demographically and educationally diverse students. In addition, outreach to universities builds relationships and awareness about the Laboratory's opportunities and capabilities in the challenging field of data science. LLNL researchers interested in mentoring Data Heroes are encouraged to contact Gansemer and Herndon Ford.



Caption: Data Heroes Olivia Simpson (left) and Eliza Wells (right) listen as Paul Rosenfield describes his work with neutron detection.